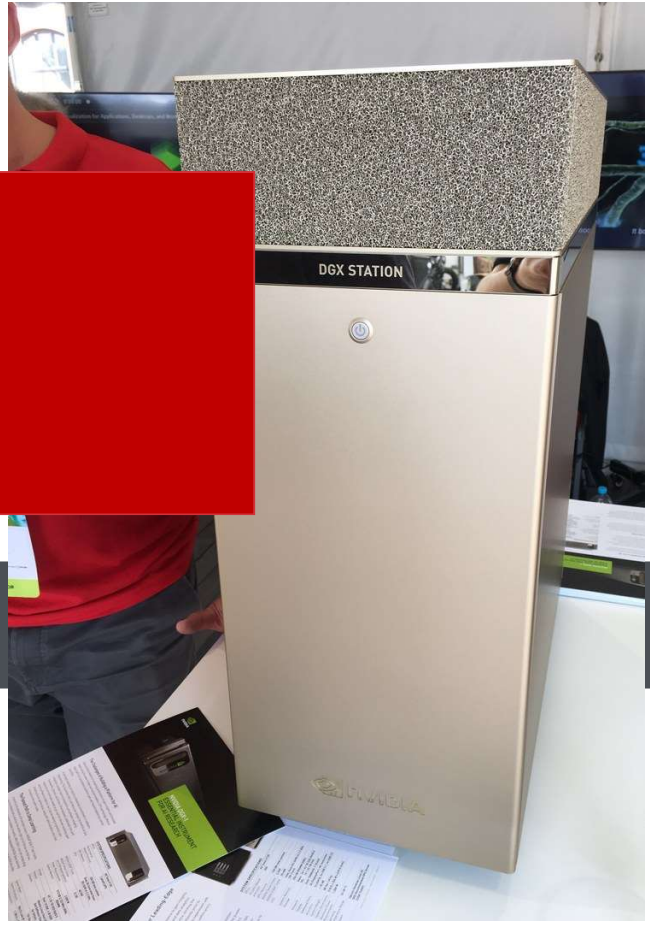
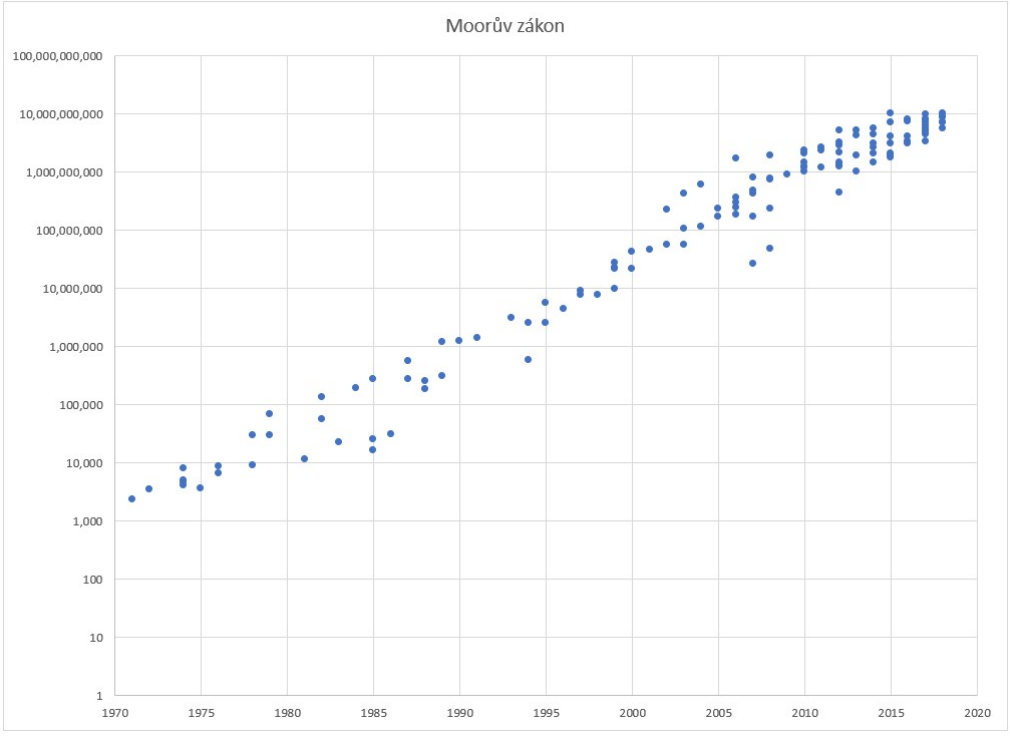
NOVINKY NVIDIA

LEDEN 2019

Petr Plodík, M Computers s.r.o.
petr.plodik@mcomputers.cz, 737 264 480

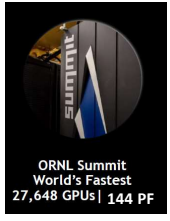


VÝKON POČÍTAČŮ



CPU:

- Intel – cluster, SMP
- AMD
- IBM Power
- ARM
- RISC-V
- …



Akcelerátory:

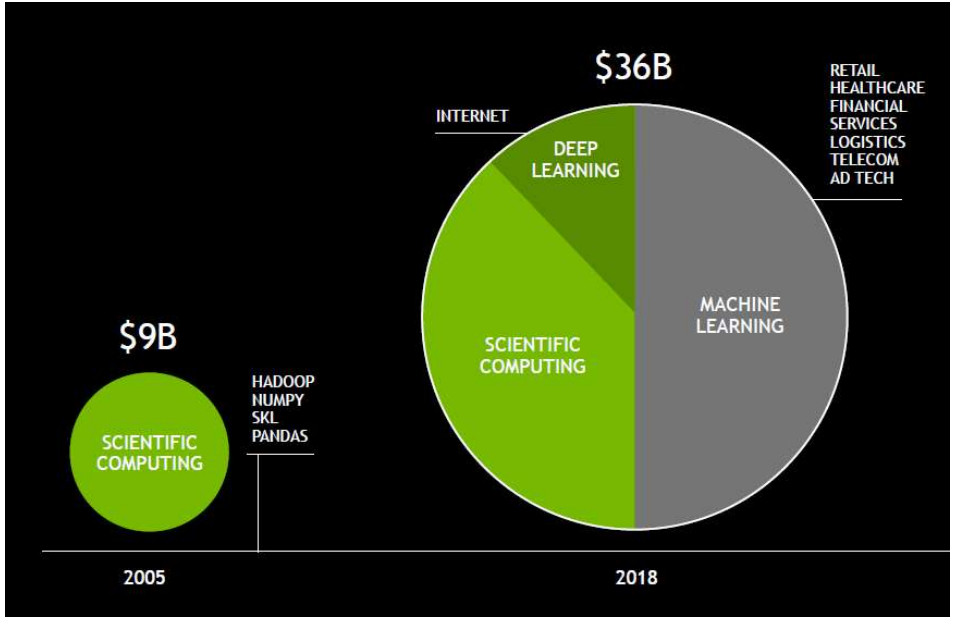
- NVIDIA Tesla
- AMD Radeon
- FPGA – Xilinx, Altera
- Google (TPU)
- Graphcore GC2 (IPU)
- Huawei Ascend
- Intel Nervana, Movidius
- NEC SX-Aurora
- …

Nové technologie?

- kvantové počítače
- 3D technologie pro CPU
- Dataflow architektury
- neuromorfické čipy
- …

KONVERGENCE HPC A AI PROSTŘEDÍ

- Provoz tradičních vědeckých **HPC aplikací** a nových **machine learning/deep learning** aplikací na jedné infrastruktuře



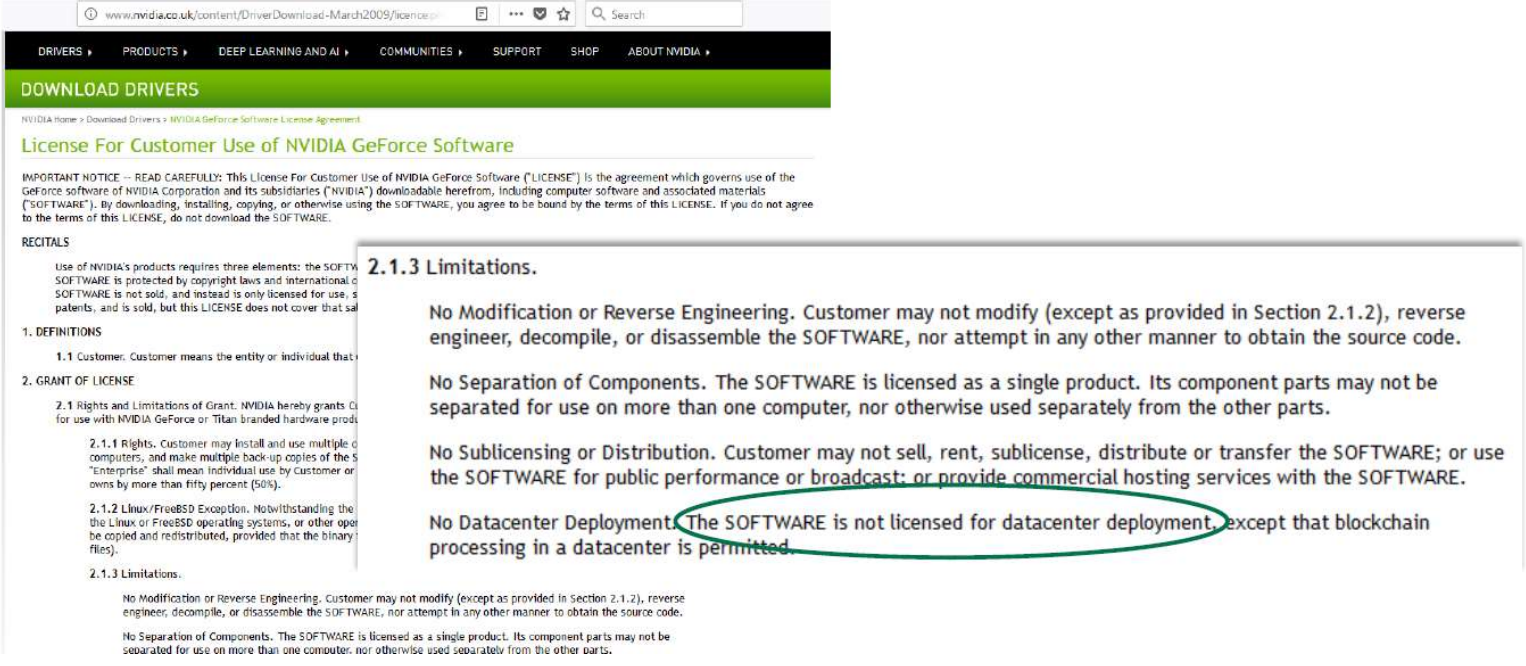
PŘEHLED AKTUÁLNÍCH NVIDIA KARET






	GeForce RTX 2080Ti	Quadro RTX 5000	Tesla T4	Tesla V100 SXM2	DGX-1	DGX-2
počet Cuda jader	4 352	3 072	2 560	5 120	40 960	81 920
počet Tensor jader	544	384	320	640	5 120	10 240
FP32 Tflops	13,4	11,2	8,1	15,7	125	250
INT8 Tflops	215,2	178,4	130	NA	NA	NA
Tensor TFlops	107,6	89,2	65	125	1 000	2 000
GP GPU Memory	11 GB	16 GB	16 GB	32 GB	256 GB	512 GB
Memory tech.	GDDR6	GDDR6	GDDR6	HMB2	HMB2	HMB2
NVLink	2-way	2-way	Ne	Ano	8-way	16-way
Max. TDP (W)	250 W	230 W	70 W	300 W	3 500 W	10 kW
určení	pouze stanice	stanice/server	server	server	Server	server
orientační EDU cena v Kč bez DPH	25 000 Kč	60 000 Kč	60 000 Kč	244 000 Kč	2M Kč + support	6,5M Kč + support

ZMĚNA LICENČNÍCH PODMÍNEK PRO OVLADAČE NA NVIDIA GEFORCE KARTY



www.nvidia.co.uk/content/DriverDownload-March2009/licence.p

Search

DRIVERS ▸ PRODUCTS ▸ DEEP LEARNING AND AI ▸ COMMUNITIES ▸ SUPPORT SHOP ABOUT NVIDIA ▸

DOWNLOAD DRIVERS

NVIDIA Home > Download Drivers > NVIDIA GeForce Software License Agreement

License For Customer Use of NVIDIA GeForce Software

IMPORTANT NOTICE – READ CAREFULLY: This License For Customer Use of NVIDIA GeForce Software ("LICENSE") is the agreement which governs use of the GeForce software of NVIDIA Corporation and its subsidiaries ("NVIDIA") downloadable herefrom, including computer software and associated materials ("SOFTWARE"). By downloading, installing, copying, or otherwise using the SOFTWARE, you agree to be bound by the terms of this LICENSE. If you do not agree to the terms of this LICENSE, do not download the SOFTWARE.

RECITALS

Use of NVIDIA's products requires three elements: the SOFTW
SOFTWARE is protected by copyright laws and international c
SOFTWARE is not sold, and instead is only licensed for use, s
patents, and is sold, but this LICENSE does not cover that sal

1. DEFINITIONS

1.1 Customer. Customer means the entity or individual that

2. GRANT OF LICENSE

2.1 Rights and Limitations of Grant. NVIDIA hereby grants C
for use with NVIDIA GeForce or Titan branded hardware prod

2.1.1 Rights. Customer may install and use multiple c
computers, and make multiple back-up copies of the S
"Enterprise" shall mean individual use by Customer or
owns by more than fifty percent (50%).

2.1.2 Linux/FreeBSD Exception. Notwithstanding the
the Linux or FreeBSD operating systems, or other oper
be copied and redistributed, provided that the binary
files).

2.1.3 Limitations.

No Modification or Reverse Engineering. Customer may not modify (except as provided in Section 2.1.2), reverse engineer, decompile, or disassemble the SOFTWARE, nor attempt in any other manner to obtain the source code.

No Separation of Components. The SOFTWARE is licensed as a single product. Its component parts may not be separated for use on more than one computer, nor otherwise used separately from the other parts.

2.1.3 Limitations.

No Modification or Reverse Engineering. Customer may not modify (except as provided in Section 2.1.2), reverse engineer, decompile, or disassemble the SOFTWARE, nor attempt in any other manner to obtain the source code.

No Separation of Components. The SOFTWARE is licensed as a single product. Its component parts may not be separated for use on more than one computer, nor otherwise used separately from the other parts.

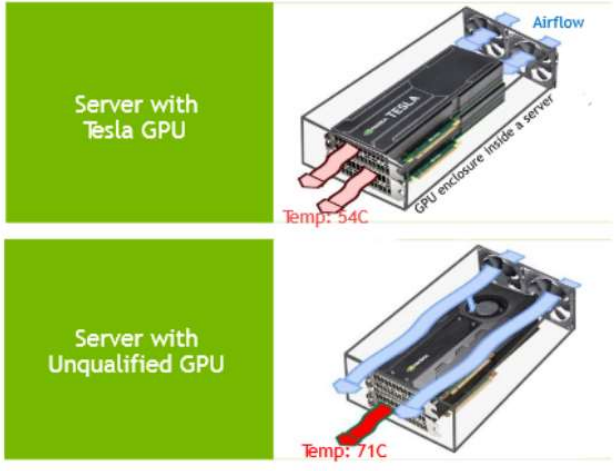
No Sublicensing or Distribution. Customer may not sell, rent, sublicense, distribute or transfer the SOFTWARE; or use the SOFTWARE for public performance or broadcast; or provide commercial hosting services with the SOFTWARE.

No Datacenter Deployment. The SOFTWARE is not licensed for datacenter deployment, except that blockchain processing in a datacenter is permitted.

VLASTNOSTI NVIDIA TESLA

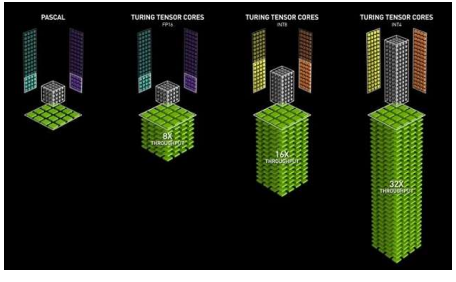
Since 2010, NVIDIA Tesla has been the standard for the datacenter.

- ✓ 24x7 Uptime & stress tested
- ✓ OEM certified and qualified
- ✓ Maximum airflow & low voltage
- ✓ Increased DP performance
- ✓ Increased memory
- ✓ Support & warranty
- ✓ Multi-GPU



NVIDIA TESLA T4

Tesla T4 – zabere jeden PCIe slot, 70W



NVIDIA TESLA T4 SPECIFICATIONS

Performance

Tesla T4 vs. Tesla P4

	FP16	INT8	INT4
Nvidia Tesla T4 (TFLOPS)	65	130	260
Nvidia Tesla P4 (TFLOPS)	5.5	22	-

Memory

CAPACITY
16 GB GDDR6

BANDWIDTH
320+ GB/s

Power
70 watts

TURING TENSOR CORES
320

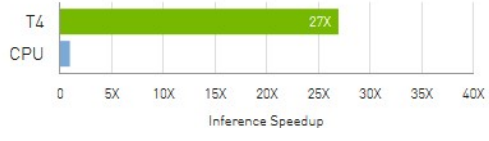
NVIDIA CUDA® CORES
2,560

SINGLE PRECISION PERFORMANCE (FP32)
8.1 TFLOPS

MIXED PRECISION (FP16/FP32)
65 FP16 TFLOPS

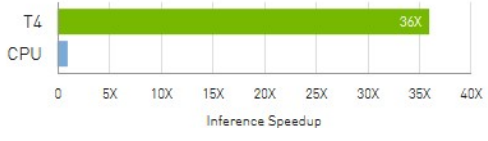
INT8 PRECISION
130 INT8 TOPS

Resnet50



DeepSpeech2

GNMT



NVIDIA DGX SYSTÉMY

https://www.mcomputers.cz/nvidia-dgx-systemy/

Parametr	DGX-2	DGX-1	DGX Station
GPUs	16× NVIDIA Tesla V100 32GB	8× NVIDIA Tesla V100 32GB	4× NVIDIA Tesla V100 32GB
Výkon (tensor operace)	2 PetaFLOPS	1 PetaFLOPS	0,5 PetaFLOPS
GPU paměť	512 GB celkem	256 GB celkem	128 GB celkem
CPU	2× Platinum 8168, 2.7 GHz (24 jader)	2× E5-2698 v4 2.2GHz (20 jader)	E5-2698 v4 2.2GHz (20 jader)
NVIDIA CUDA cores	81 920	40 960	20 480
NVIDIA Tensor cores	10 240	5 120	2 560
Propojení GPU karet	NVSwitch, non-blocking, 2,4TB/s	NVLink, hypercube topologie	NVLink
RAM	1,5 TB	512 GB	256 GB
HDD	2× 960GB NVME SSD, 8× 3.84TB NVME SSD	4× 1,92TB SSD	4× 1,92TB SSD
Network	2× 10/25GbE, 8× 100Gb EDR Infiniband/Ethernet	2× 10GbE, 4× 100Gb EDR Infiniband/Ethernet	2× 10GbE
Maximální příkon	10 kW	3 500 W	1 500 W
Provedení	rack, 10U	rack, 3U	tower, vodní chlazení CPU a GPU

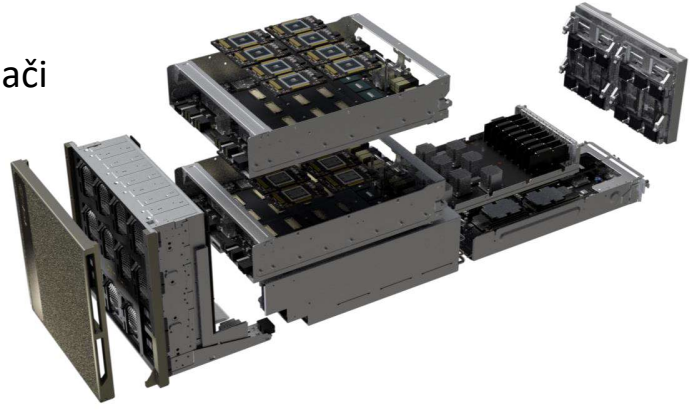
DGX-2, HGX-2 A NVSWITCH

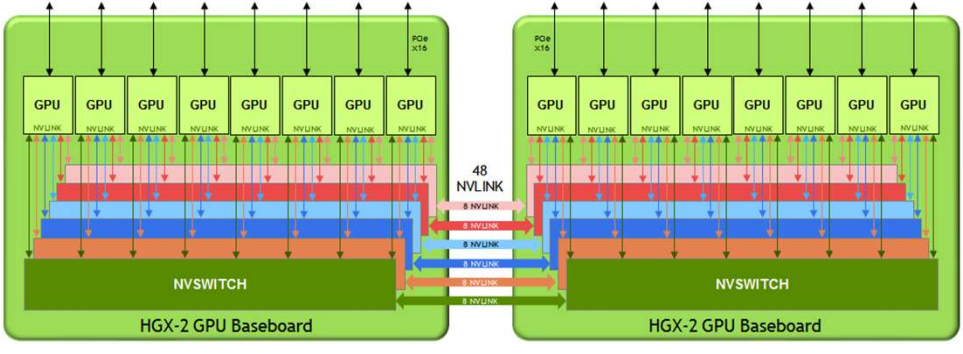
16× Tesla V100 32GB propojených NVSwitch přepínači
2 petaFLOPS deep learning výkon

sdílená 512GB HBM2 GPU paměť

300GB/s GPU-GPU propustnost

2.4TB/s celková cross section propustnost





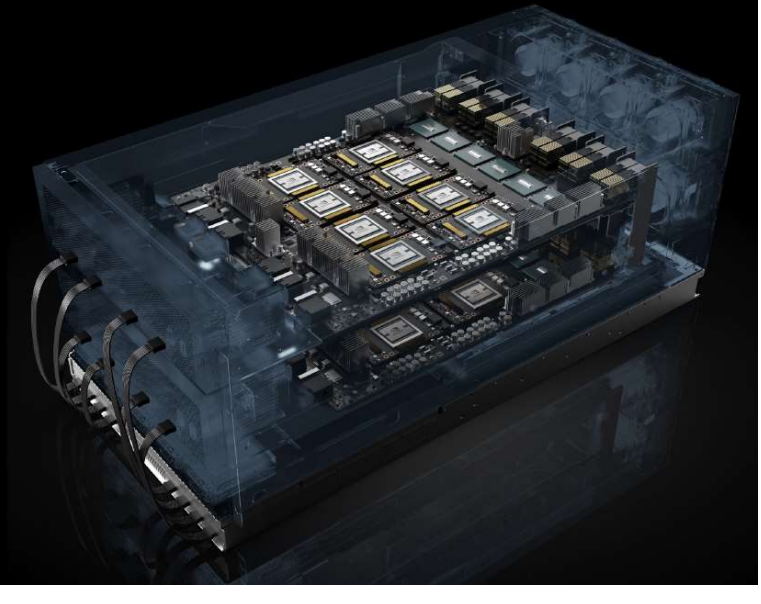
NVIDIA NVSwitch

TESLA HGX-2

Fusing HPC and AI into One Unified Computing Architecture

Multi-precision Computing
2 PFLOPS AI | 250 TFLOPS FP32
125 TFLOPS FP64
16 Tesla V100 GPUs
2.4 TB/s NVLink Bisection Bandwidth

16TB/s Memory Bandwidth | 0.5TB Memory
Flexible Platform Solution



NVIDIA DGX REFERENČNÍ ARCHITEKTURY

NVIDIA DGX systémy + networking + disková pole

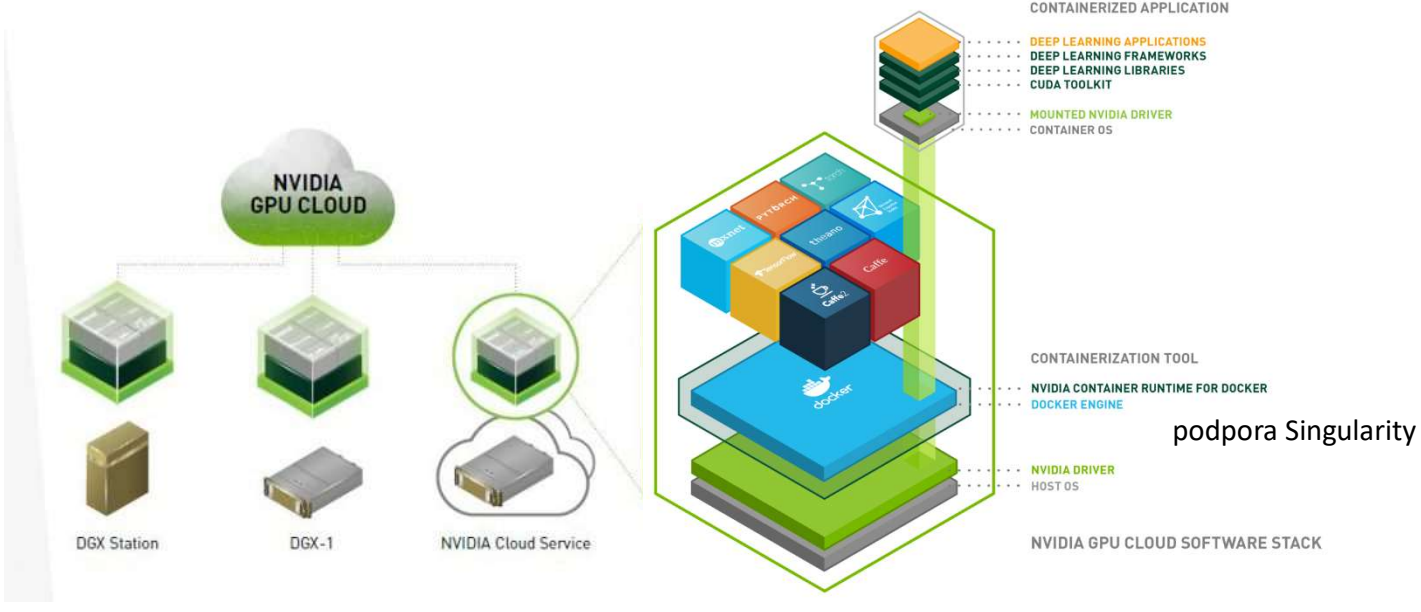
NVIDIA má obecnou architekturu NVIDIA DGX POD Reference Architecture

Z ní vychází návrhy jednotlivých storage vendorů:

- NetApp ONTAP AI Reference Architecture + Cisco
- DDN A3I Reference Architecture + Mellanox
- Pure Storage AIRI Reference Architecture + Arista
- IBM Spectrum AI Reference Architecture

NVIDIA GPU CLOUD/CONTAINERS (NGC)

NVIDIA docker image repository: https://ngc.nvidia.com



podpora Singularity

NVIDIA DGX OPERATING SYSTEM

Release 4.0 https://docs.nvidia.com/dgx/dgx-os-server-release-notes/index.html

•Highlights
- •First release to support the NVIDIA DGX-2 System
- •Ubuntu 18.04 LTS
- •NVIDIA GPU Driver Release 410
  - •Supports CUDA 10.0
- •Docker CE and the NVIDIA Container Runtime for Docker are pre-installed, and the docker daemon automatically launched.
- •New NVIDIA System Management (NVSM) health monitoring software framework

Replaces nvsysinfo and nvhealth.

•See the following for additional information and known issues for the latest version.
- •***NEW***DGX OS Server v4.0.4 Release Notes v02
- •DGX OS Server v4.0.3 Release Notes v03(DGX-2 System only)
- •DGX OS Server v4.0.2 Release Notes v01 (DGX-2 System only)

Release 3.1

•Highlights
- •Ubuntu 16.04 LTS
  - •Initialization daemon changed from **Upstart** to **systemd**.
  - •Updated network interface naming policy.

Policy now uses predictable names, rather than the native naming scheme used in previous releases.
- NVIDIA GPU Driver Release 384
  - •Supports the NVIDIA Tesla™ V100 GPUs.
  - •Supports CUDA 9.0 .
- •CUDA drivers and diagnostic packages updated to Release 384.
- •Mellanox drivers updated to 4.0.
- •Docker CE and nvidia-docker are pre-installed, and the docker daemon automatically launched.

•See the following for additional information, known issues, and update instructions for specific versions.
- •*** NEW***DGX OS Server v3.1.7 Release Notes v03
- •DGX OS Server v3.1.6 Release Notes
- •DGX OS Server v3.1.4 Release Notes
- •DGX OS Server v3.1.2 Release Notes
- •DGX OS Server v3.1.1 Release Notes

VERSION HISTORY

Version 4.0.4

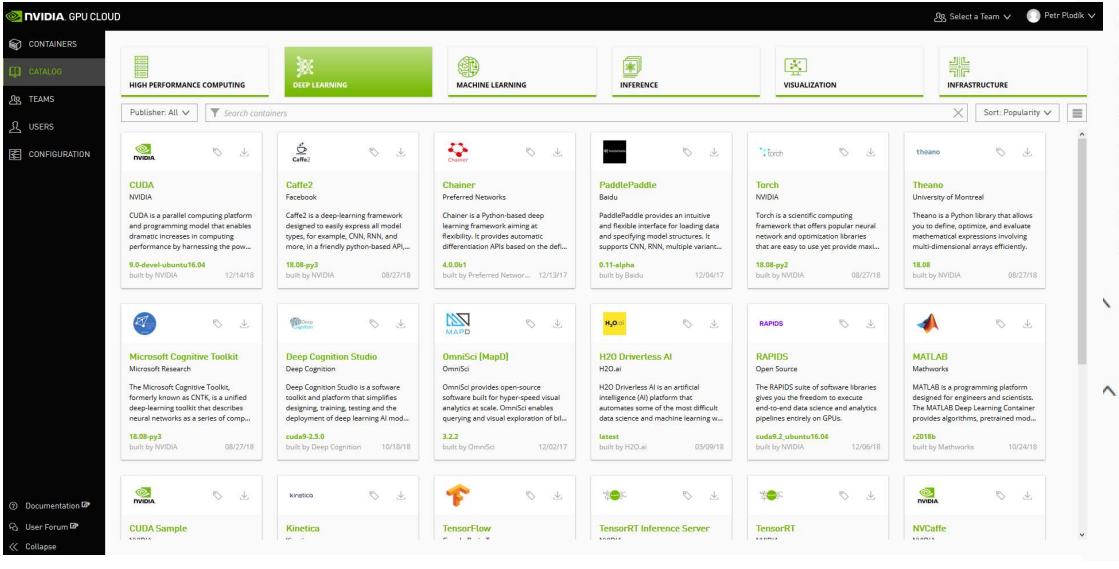
- Added support for NVIDIA DGX-1 systems.
  Sets default Ubuntu IO scheduler from **CFQ** to **deadline**.
- Updated NVIDIA GPU driver to version 410.79.

Version 4.0.3

- Updated NVIDIA GPU driver to version 410.72.
- Updated other software components.
  - DCGM updated to version 1.5.3
  - Docker updated to version 18.06.1-ce
  - NVSM components updated to version 18.10 (See DGX OS Server Software Content for details)
- Updated KVM² Software.
  - KVM software (dgx-kvm-sw) updated to version 18.10.2.
  - KVM image (dgx-kvm-image) updated to version 4-0-3.
  - Added FS-Cache support for guest VMs.
  - Added multi-queue support for logical drives.
  - Added multi-queue support for virtual networking.
  - Added NUMA tuning.
  - Added CPU tuning (emulatorpin).

NVIDIA GPU CLOUD (NGC)

NVIDIA docker image repository: https://ngc.nvidia.com



nvidia/k8s

- cuda-sample
- dcgm-exporter
- device-plugin
- etcd-amd64
- k8s-dns-dnsmasq-nann...
- k8s-dns-kube-dns-amd64
- k8s-dns-sidecar-amd64
- kube-aggregator-amd64
- kube-apiserver-amd64
- kube-controller-manag...
- kube-proxy-amd64
- kube-scheduler-amd64

nvidia/rapidsai

- rapidsai

nvidia-hpcvis

- index
- paraview-holodeck
- paraview-index
- paraview-optix

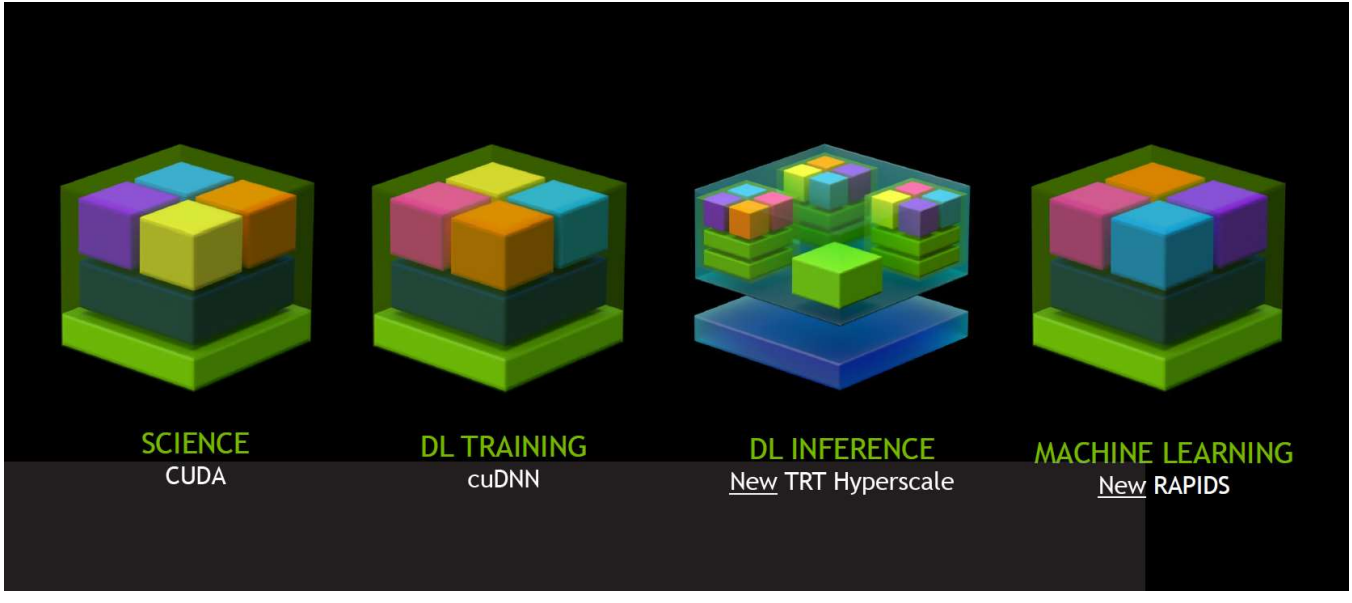
hpc

- bigdft
- candle
- chroma
- gamess
- gromacs
- lammps
- lattice-microbes
- milc
- namd
- pgi-compilers
- picongpu
- relion
- vmd

partners

- chainer
- h2oai-driverless
- kinetica
- mapd
- matlab
- paddlepaddle

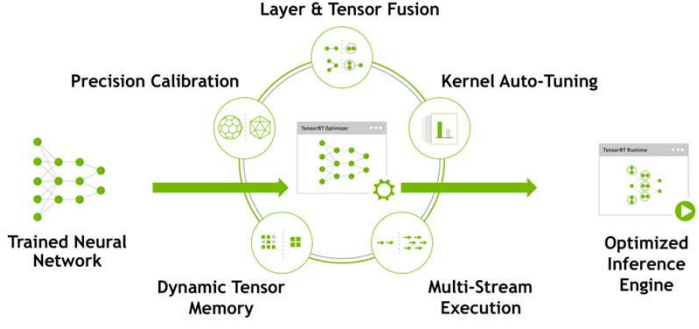
NVIDIA FULL SOFTWARE STACK



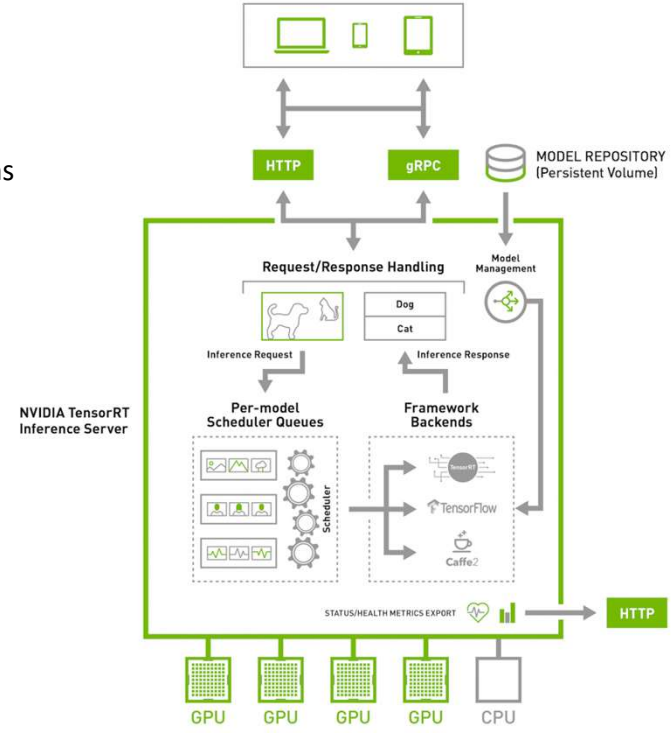
NVIDIA TENSORRT 5

https://developer.nvidia.com/tensorrt

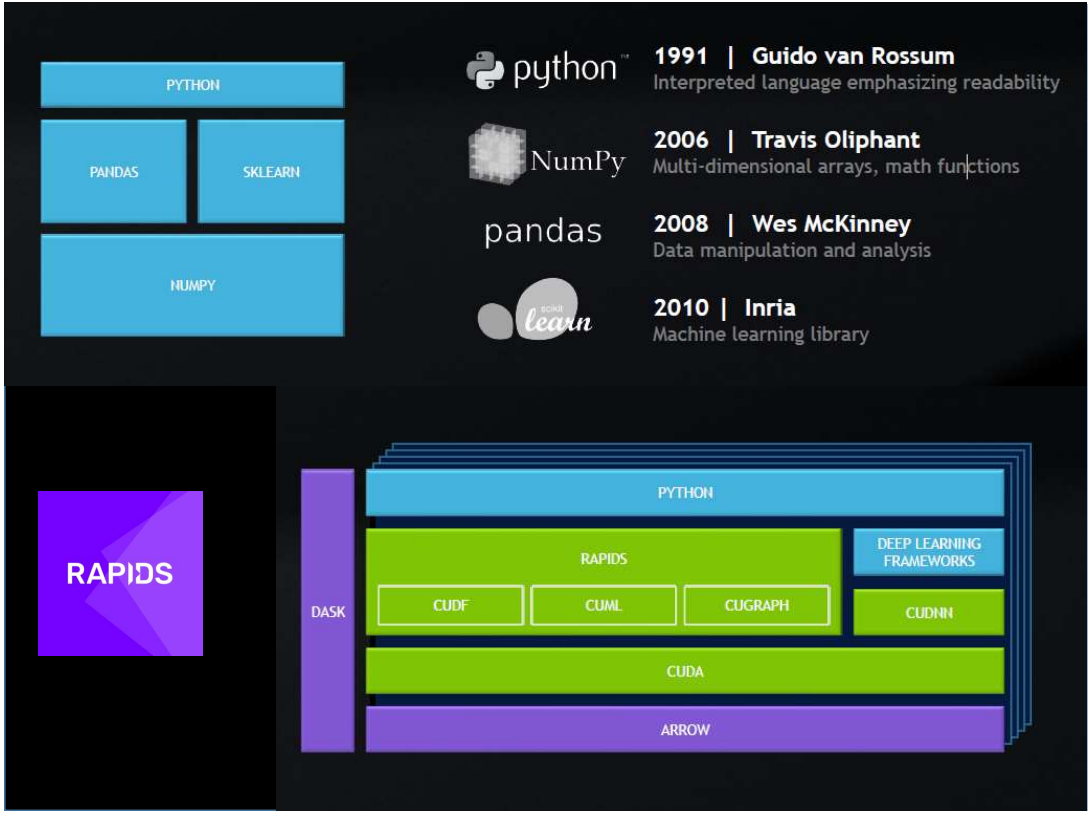
- Speed up inference by 40× over CPUs for models such as translation using mixed precision on Turing Tensor Cores
- Optimize inference models with new INT8 APIs and optimizations
- Deploy applications to Xavier-based NVIDIA Drive platforms and the NVIDIA DLA accelerator (FP16 only)
- **NVIDIA TRT Hyperscale** -- Docker and Kubernetes



TensorRT Inference Server



NVIDIA RAPIDS



NVIDIA RAPIDS

https://developer.nvidia.com/rapids

DataFrame - cuDF - This is a GPU accelerated DataFrame-manipulation library based on GPU Apache Arrow. It's designed to enable data wrangling data for model training. The Python bindings of the core-accelerated, low-level CUDA C++ kernels mirror the pandas API for seamless onboarding and transition from pandas.

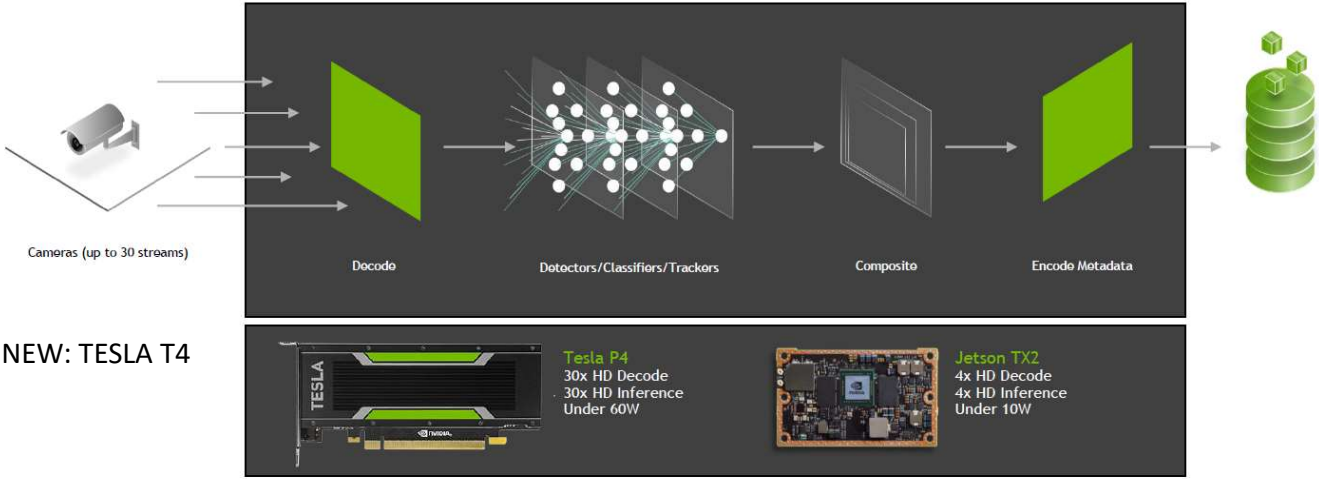
Machine Learning Libraries - cuML - This collection of GPU-accelerated machine learning libraries will eventually provide GPU versions of all machine learning algorithms available in Scikit-Learn.

Graph Analytics Libraries - cuGRAPH - This collection of graph analytics libraries that seamlessly integrates into the RAPIDS data science software suite.

Deep Learning Libraries - RAPIDS provides native array_interface support. This means data stored in Apache Arrow can be seamlessly pushed to deep learning frameworks that accept array_interface such as PyTorch and Chainer.

Visualization Libraries - Coming soon. RAPIDS will include tightly integrated data visualization libraries based on Apache Arrow. Native GPU in-memory data format provides high-performance, high-FPS data visualization, even with very large datasets.

DEEPSTREAM
UP TO 30 HD STREAMS AT 30 FPS



NEW: TESLA T4

https://developer.nvidia.com/deepstream-sdk

ZKUŠENOSTI Z TESTOVÁNÍ NVIDIA DGX STATION

NVIDIA DGX Station: 1× Intel Xeon, 4× NVIDIA Tesla V100 32GB PCIe propojené NVLinkem

+ snadná instalace – předinstalovaná image s NVIDIA OS: Ubuntu + Docker

+ předpřipravené docker/singularity kontejnery v NVIDIA NGC

+ tichý provoz díky vodnímu chlazení CPU a GPU karet

vs.

nevyužití potenciálu výpočetních karet NVIDIA Tesla V100:

- aplikace nedokáže využít tensor jádra GPU karty
- nevyužití více GPU karet v jednom serveru a NVLink rozhraní
- aplikace nedokázali využít celou GPU paměť (menší modely)

NVIDIA EDU PROGRAM

Speciální ceny na vybrané karty a DGX systémy, vyhlašované každý kvartál.

Např. sleva 1 600 USD na NVIDIA Tesla V100 32GB

30% sleva na NVIDIA DGX systémy, orientační ceny:

- NVIDIA DGX Station: 1,1M Kč bez DPH + podpora
- NVIDIA DGX-1: 2,4M Kč bez DPH + podpora
- NVIDIA DGX-2: 6,5M Kč bez DPH + podpora

Sleva na NVIDIA Jetson/AGX development kity

JETSON TX2/TX2i

Možnosti testování karet – NVIDIA Test Drive:

- NVIDIA DGX Station
- NVIDIA V100
- NVIDIA T4

JETSON XAVIER: AGX

NVIDIA GTC CONFERENCE

https://www.nvidia.com/en-eu/gtc/

EXPERIENCE THE POWER OF GTC EUROPE

NVIDIA's GPU Technology Conference (GTC) Europe is part of the largest global series of events focused on artificial intelligence and its applications across many important fields.

Join us in Munich and discover the latest breakthroughs in autonomous vehicles, high performance computing, healthcare, big data, and more.

CONNECT

Network with experts from NVIDIA and other leading organisations.

LEARN

Learn how to solve challenging problems with deep learning and accelerated computing through hands-on training.

DISCOVER

Discover the latest breakthroughs in a wide range of fields such as deep learning, self-driving vehicles, HPC, virtual reality, and more.

INNOVATE

Hear about disruptive innovations as startups and researchers present their work.

Pro získání **20% slevy** můžete při registraci na konferenci využít náš promo kód

„Naše řešení vás budou bavit“

www.mcomputers.cz